# New kurtosis optimization schemes
# for MISO equalization

Marc Castella\*, *Member, IEEE*, and Eric Moreau<sup>†</sup>, *Senior Member, IEEE*

\* Institut Télécom; Télécom Sudparis; UMR-CNRS 5157
9 rue Charles Fourier, 91011 Évry Cedex, France
Email: marc.castella@it-sudparis.eu / Fax: (+33) 1 60 76 44 33 / Tel: (+33) 1 60 76 41 71

<sup>†</sup> University of Sud Toulon Var, ISITV, LSEET UMR-CNRS 6017
av. G. Pompidou, BP56, F-83162 La Valette du Var Cedex, France
Email: moreau@univ-tln.fr / Fax: (+33) 4 94 14 24 48 / Tel: (+33) 4 94 14 25 86

*Abstract*—This paper deals with efficient optimization of cumulant based contrast functions. Such a problem occurs in the blind source separation framework, where contrast functions are criteria to be maximized in order to retrieve the sources. More precisely, we focus on the extraction of one source signal and our method applies in deflation approaches, where the sources are extracted one by one.

We propose new methods to maximize the kurtosis contrast function. These methods are intermediate between a gradient and an iterative "fixed-point" optimization of so-called reference contrasts. They rely on iterative updates of the parameters which monotonically increase the contrast function value: we point out the strong similarity with the EM (Expectation-Maximization) method and with recent generalizations referred to as MM (Minimization-Maximization). We also prove the global convergence of the algorithm to a stationary point. Simulations confirm the convergence of our methods to a separating solution. They also show experimentally that our methods have a much lower computational cost than former classical optimization methods. Finally, simulations suggest that the methods remain valid under weaker conditions than those required for proving convergence.

*Index Terms*—contrast function, reference system, quadratic optimization, higher-order statistics, global convergence

## I. INTRODUCTION

### A. Generalities

The general problem of signal restoration has been given a constant interest in the signal processing literature. In particular, the source separation problem has received a considerable attention due to its wide range of applications such as communications, biology, seismology, radar, . . . . In this contribution, we assume that no information on the mixing system is used for the separation, but only the observations may be used: this context is referred to as *blind*. The problem of (blind) equalization in digital communications is a tightly connected one [1]. It is known that such techniques can be used for throughput increase.

In a multi-input/multi-output (MIMO) convolutive context, the problem of blind source separation has found interesting solutions through the optimization of so-called contrast functions [2]. Among the possible approaches, the source signals can be either separated simultaneously [3]–[6] or extracted

one by one by optimizing for each a multi-input/single-output (MISO) separating criterion. In this paper, we consider the latter approach, which is often combined with a deflation procedure to extract all the sources [7]–[9].

Many separation criteria rely on higher-order statistics (e.g. the kurtosis contrast [8], [9]) or can be linked to higher-order statistics (e.g. the constant modulus contrast function [10]). These criteria are known to provide good results. On the other hand, contrast functions referred to as "reference-based" have been recently proposed [11], [12]. They are particularly appealing because the corresponding maximization problem is quadratic with respect to the searched parameters. Taking advantage of this quadratic feature, a maximization algorithm based on singular value decomposition (SVD) has been proposed [11], [13] and was shown to be significantly quicker than other maximization algorithms. In the case where the "reference signal" is close to an actual source, the SVD based method is very efficient for the extraction of this source. However, the method generally requires an additional "fixed-point" like iteration to improve the separation quality [11]. Also, the SVD based optimization is very sensitive to a rank estimation and thus it is not appropriate to use it within a deflation procedure. In this context, it should be replaced by a gradient optimization [14].

In this paper, we propose a new family of algorithms to maximize a kurtosis based contrast function. All optimization methods derive from a basic algorithm which is first presented and whose global convergence to a stationary point is proved. Then, the necessity of a renormalization step is discussed, which yields the first algorithm applicable in practice. Finally, the method is extended by introducing two different number of iteration parameters. Depending on these parameters, a trade-off can be adjusted between performance and speed of the optimization method. The proposed method is inspired from the gradient optimization of a "reference-based" contrast function, which has been proposed in [14]. On the other hand, the gradient optimization of the kurtosis contrast function [8], [9] has been used for long. Our method is different from both gradient optimization methods, but it can be interpreted as being intermediate between the two: indeed, a "reference-based" iterative optimization is regularly interrupted and started again

with the initialization point newly obtained. In this paper, the main novelties which did not appear in [14], [15] are the following ones:

- a detailed proof of the global convergence of the optimization algorithms,
- a link with EM (Expectation-Maximization) and MM (Minimization-Maximization) algorithms, which justifies the convergence of the "fixed-point" method introduced in [11],
- the possibility in the new method to adjust two parameters for improved performance.

Section II describes the considered model and assumptions. In Section III, the contrast function and the separating method are given. The ideas on which the optimization method relies are explained in Section IV whereas the technical details and the complete proof are relegated to Appendices A and B. Section IV-B also establishes the connection with other optimization methods. Section V is devoted to improvements and practical implementation details: the practical algorithms (**Alg.2**) is described in this section. Finally, Section VI presents simulations results and Section VII concludes the paper.

### B. Notations

In the whole paper, $n$ stands for a generic integer ($n \in \mathbb{Z}$) and the cumulant of a set of random variables is denoted $\mathrm{Cum}\{.\}$. The signals can be complex- or real-valued.

For any jointly stationary signals $z(n)$ and $y(n)$, we set $\mathrm{C}\{y\} \triangleq \mathrm{Cum}\{y(n), y(n)^*, y(n), y(n)^*\}$ and $\mathcal{C}_z\{y\} \triangleq \mathrm{Cum}\{y(n), y(n)^*, z(n), z(n)^*\}$. For the sources, we will adopt also the specific notation $\kappa_i \triangleq \mathrm{Cum}\{s_i(n), s_i(n)^*, s_i(n), s_i(n)^*\}$. Finally, brackets $\{.\}$ will be used as shorthand notations to denote linear time invariant (LTI) filters and $\star$ denotes filter convolution. The output of any filter will be denoted such as in Equation (1) in next Section to avoid writing the summation.

## II. MODEL AND ASSUMPTIONS

### A. Convolutive mixture

We consider an observed $Q$-dimensional ($Q \in \mathbb{N}, Q \geq 2$) discrete-time signal. Its $n$th sample is denoted by the column vector $\mathbf{x}(n)$. Assuming a noise-free linear convolutive model, the observation $\mathbf{x}(n)$ results from a linear time invariant (LTI) multichannel system described by the input-output relation:

$$\mathbf{x}(n) = \sum_{p \in \mathbb{Z}} \mathbf{M}(p)\mathbf{s}(n - p) \triangleq \{\mathbf{M}\}\mathbf{s}(n). \tag{1}$$

In the above equation, $\mathbf{M}(n)$ represents a $Q \times N$ matrix corresponding to the impulse response of the LTI mixing system. For simplicity, we will use the shorthand notations $\{\mathbf{M}\}$ to denote this MIMO LTI filter and $\{\mathbf{M}\}\mathbf{s}(n)$ to denote the above sum. The vector of source signals (or *sources*) $\mathbf{s}(n)$ is an $N$-dimensional ($N \in \mathbb{N}^*$) unknown and unobserved column vector. The objective is to restore the sources *blindly*, that is from the only use of the observations.

We assume that the LTI mixing system admits a left LTI inverse $\{\mathbf{W}\}$ such that $\mathbf{y}(n) \triangleq \{\mathbf{W}\}\mathbf{x}(n) \triangleq \sum_{p \in \mathbb{Z}} \mathbf{W}(p)\mathbf{x}(n-$

$p)$ corresponds to the sources. In other words, the global system $\{\mathbf{W} \star \mathbf{M}\}$ is the identity up to some possible ambiguities to be specified next.

More specifically, our approach is an iterative one and we will focus on the extraction of a single source. The extraction of all sources can then be performed through a deflation procedure [8], [16], [17]. In this MISO context, the considered problem consists in estimating one row of $\{\mathbf{W}\}$. The corresponding row is a $1 \times Q$ LTI vector filter $\{\mathbf{w}\}$, called equalizer. Its impulse response is $\mathbf{w}(n)$ and the output of the separation procedure is the scalar signal

$$y(n) = \sum_{p \in \mathbb{Z}} \mathbf{w}(p)\mathbf{x}(n - p). \tag{2}$$

Ideally, $y(n)$ should restore one of the components $s_i(n), i \in \{1, \ldots, N\}$, of the source vector: in other words, under successful separation, $y(n)$ should be equal to a scalar filtering of one of the scalar source signals $s_i(n), i \in \{1, \ldots, N\}$. It is also well-known that the remaining scalar filtering ambiguity reduces to a delay and a scalar factor in the case of i.i.d. sources (see [2], [8], [11] for more details). In order to obtain tractable expressions, we define the corresponding $1 \times N$ global LTI vector filter $\{\mathbf{g}\}$ by its impulse response $\mathbf{g}(n) \triangleq \sum_{p \in \mathbb{Z}} \mathbf{w}(p)\mathbf{M}(n - p)$. Then we have

$$y(n) = \sum_{p \in \mathbb{Z}} \mathbf{g}(n - p)\mathbf{s}(p) \triangleq \{\mathbf{g}\}\mathbf{s}(n) . \tag{3}$$

The notations are summed up in Figure 1. Note that in simulations, the filters $\{\mathbf{w}\}$ and $\{\mathbf{v}\}$ appearing in Figure 1 will be chosen as FIR of length $D$ (see Section V-B).
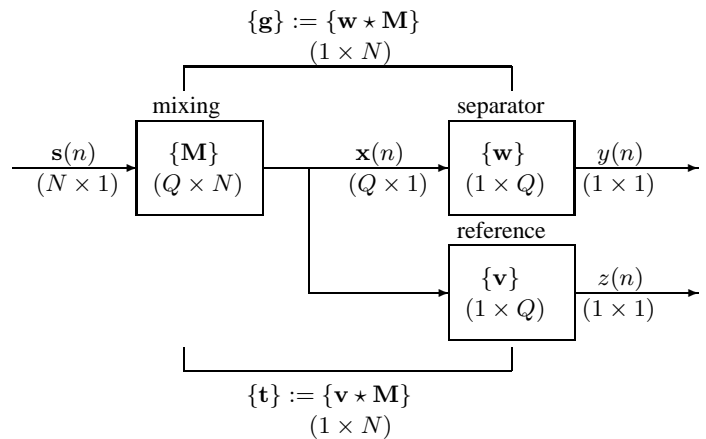


Fig. 1. System and notations summary

### B. Assumptions on the sources

To be able to carry out the estimation blindly, some assumptions about the source signals are required. In this paper, we will adopt the following assumptions:

A1. For all $i$, the source sequence $s_i(n)$ is stationary, zero-mean and with unit variance. In addition, the fourth-order cumulants $\kappa_i \triangleq \mathrm{Cum}\{s_i(n), s_i(n)^*, s_i(n), s_i(n)^*\}$ exist and are assumed to be nonzero.

A2. The source vector processes $s_i(n), i \in \{1, \ldots, N\}$ are statistically mutually independent.

## III. SEPARATION METHOD

### A. Contrast function

It has been proved in [8], [9] that the criterion $\left|\frac{\mathrm{C}\{y\}}{\mathbb{E}\{|y|^2\}^2}\right|$ is a contrast function: this means by definition that when maximized with respect to the separating filter, this criterion leads to the separation of one source signal. In this paper, we propose new methods to maximize this kurtosis contrast function.

Our approach is tightly related to the concept of "reference-based" contrast functions [11]. We introduce a so-called "reference signal" which is the output of a $1 \times Q$ LTI "reference filter" denoted by $\{\mathbf{v}\}$ (see Figure 1). The "reference filter" $\{\mathbf{v}\}$ is driven by the observations $\mathbf{x}(n)$. Ideally, the "reference filter" should be as close as possible to a separating filter. Similarly to Equations (2)-(3), we introduce the corresponding $1 \times N$ global LTI filter $\{\mathbf{t}\} = \{\mathbf{v} \star \mathbf{M}\}$ and the reference signal then reads:

$$z(n) = \{\mathbf{t}\}\mathbf{s}(n) = \sum_{i=1}^{N} \sum_{p \in \mathbb{Z}} t_i(p) s_i(n-p) \quad \text{or:}$$

$$z(n) = \{\mathbf{v}\}\mathbf{x}(n) = \sum_{i=1}^{Q} \sum_{p \in \mathbb{Z}} v_i(p) x_i(n-p) \quad (4)$$

In our previous work [11], we have given the conditions required by the reference signal to obtain new contrast functions. These conditions are very weak and practically always satisfied. In our new point of view, the reference signal is artificially introduced in an algorithm for the purpose of facilitating the maximization of the kurtosis contrast function. It hence appears as an efficient way of exploiting the properties of the criterion to be maximized. This is in contrast to our former works, where the reference signal could be be interpreted as a prior knowledge on the source to be extracted. Based on the latter idea, we have proposed a "fixed-point" algorithm, where the reference is successively updated by the new source estimate. We will see how this method re-enters in our general point of view. In particular, we prove the global convergence of our algorithm to a stationary point, which implies that conditions on the reference signal are no longer required. In passing, we better understand the behavior and properties of cumulant based contrast functions.

Let us introduce the following criteria:

$$\mathcal{J}(\mathbf{w}) = \left|\frac{\mathrm{C}\{y\}}{\mathbb{E}\{|y(n)|^2\}^2}\right|^2 \quad \text{and:}$$

$$\mathcal{I}(\mathbf{w}, \mathbf{v}) = \left|\frac{\mathcal{C}_z\{y\}}{\mathbb{E}\{|y(n)|^2\}\mathbb{E}\{|z(n)|^2\}}\right|^2 \quad (5)$$

The criterion $\mathcal{J}$ is the well-known kurtosis contrast function [8], [9] whereas $\mathcal{I}$ corresponds to so-called "reference contrasts" which have been introduced [11]. Obviously, we have the symmetry $\mathcal{I}(\mathbf{w}, \mathbf{v}) = \mathcal{I}(\mathbf{v}, \mathbf{w})$ and both criteria are linked by the relation $\mathcal{J}(\mathbf{w}) = \mathcal{I}(\mathbf{w}, \mathbf{w})$. Note in passing that the criteria are normalized so as to satisfy $\mathcal{I}(\lambda\mathbf{w}, \mathbf{v}) = \mathcal{I}(\mathbf{w}, \mathbf{v})$ and $\mathcal{J}(\lambda\mathbf{w}) = \mathcal{J}(\mathbf{w})$.

We introduce a gradient operator $\nabla$ and partial gradient operators $\nabla_1$ (resp. $\nabla_2$) with respect to the first (resp. second)

argument. More precisely, $\nabla\mathcal{J}(\mathbf{w})$ is the vector composed of all partial derivatives of $\mathcal{J}(\mathbf{w})$, whereas $\nabla_1\mathcal{I}(\mathbf{w}, \mathbf{v})$ (resp. $\nabla_2\mathcal{I}(\mathbf{w}, \mathbf{v})$) is the vector of partial derivatives of $\mathcal{I}(\mathbf{w}, \mathbf{v})$ w.r.t. $\mathbf{w}$ (resp. $\mathbf{v}$). It follows from the symmetry of $\mathcal{I}$ and from the chain rule that:

$$\nabla_1\mathcal{I}(\mathbf{w}, \mathbf{v}) = \nabla_2\mathcal{I}(\mathbf{v}, \mathbf{w}) \qquad \text{and:}$$
$$\nabla\mathcal{J}(\mathbf{w}) = 2\nabla_1\mathcal{I}(\mathbf{w}, \mathbf{w}) = 2\nabla_2\mathcal{I}(\mathbf{w}, \mathbf{w}) \quad (6)$$

An algorithm is proposed in the next section, which takes advantage of the properties of $\mathcal{I}$ and $\mathcal{J}$ in order to efficiently maximize the kurtosis criterion $\mathcal{J}$. One should note in passing that it is equivalent to maximize the square modulus or the modulus of the criterion. In (5), the square modulus has been introduced to ensure the differentiability of $\mathcal{I}$ and $\mathcal{J}$.

### B. Optimization algorithm

We now introduce our new algorithm for maximization of the kurtosis based contrast $\mathcal{J}$. We first give the basic and simplest version of the algorithm and give its convergence properties. Further refinements and practical considerations will be addressed in Sections IV-A and V.

<div style="border:1px solid black; padding:8px;">

**Algorithm Alg.0**

- Initialize $\mathbf{v}_0$ and the corresponding reference signal $z_0(n) = \{\mathbf{v}_0\}\mathbf{x}(n)$.
- For $k = 0, 1, 2, \ldots, k_{\max} - 1$, repeat *(M0-U)*:

  *(M0)* $\quad *$ Set $\mathbf{d}_k = \nabla_1\mathcal{I}(\mathbf{v}_k, \mathbf{v}_k)$,
  $\quad\quad\quad *$ One-dimensional optimization:
  $\quad\quad\quad\quad \alpha_k = \arg\max_\alpha \mathcal{I}(\mathbf{v}_k + \alpha\mathbf{d}_k, \mathbf{v}_k)$.

  *(U)* Update: $\mathbf{v}_{k+1} \leftarrow \mathbf{v}_k + \alpha_k\mathbf{d}_k$.

</div>

The proposed algorithm shows strong similarities with a steepest ascent algorithm. Indeed, according to (6), the algorithm moves from one point to another following the gradient direction of the criterion $\mathcal{J}$. The noticeable difference is that during the one-dimensional optimization, the considered criterion is $\mathcal{I}(\mathbf{w}, \mathbf{v}_k)$ with $\mathbf{v}_k$ fixed instead of $\mathcal{J}(\mathbf{w})$. The convergence of the algorithm to a satisfying solution point hence requires justification. The following assumption is required:

**A3.** For all $i \in \{1, \ldots, N\}$, the sources processes $s_i(n)$ are i.i.d. and the fourth order cumulants $\kappa_i$ have the same sign for all $i \in \{1, \ldots, N\}$.

Now, we can state the following proposition which is proved in Appendix A:

*Proposition 1:* Assume that the sequence $(\mathbf{v}_k)_{k \in \mathbb{N}}$ is obtained according to algorithm **Alg.0** with $k_{\max}$ infinite and that all $\mathbf{v}_k, k \in \mathbb{N}$ are contained in a compact set. Then, under assumptions A1-A3, any convergent sub-sequence of $(\mathbf{v}_k)_{k \in \mathbb{N}}$ converges to a point $\mathbf{v}^*$ such that $\nabla\mathcal{J}(\mathbf{v}^*) = 0$.

### C. Comments

Proposition 1 asserts the global convergence of the algorithm to a stationary point of the criterion $\mathcal{J}$. However, similarly to the behavior of a steepest ascent algorithm, filters corresponding to minima of the criterion should never be obtained. Also, although the algorithm might always converge

to a saddle-point in theory, such a point is unstable and we did not observe any convergence to such a point in practice. It hence follows that, practically, the algorithm yields a filter $\mathbf{v}^*$ which maximizes $\mathcal{J}(\mathbf{v})$. Since it is known that any local maximum of the criterion corresponds to a separating filter [8], it follows that *the filters obtained with our algorithm are separating ones in practice*, with only theoretical exceptions. This justifies the importance of Proposition 1.

The previous result can also be understood from a different point of view: one can see that at each step, the algorithm **Alg.0** maximizes $\mathbf{w} \mapsto \mathcal{I}(\mathbf{w}, \mathbf{v}_k)$ along a gradient direction. This can be interpreted as a one-dimensional maximization of a "reference-based" contrast along the gradient direction. The proposed algorithm can thus be understood as an intermediate method lying between the following two methods:

- a gradient ascent on the kurtosis $\mathcal{J}(\mathbf{w})$: the difference is that a "reference-based" criterion is considered in **Alg.0** during each one-dimensional optimization.
- a gradient ascent on a "reference-based" contrast $\mathcal{I}(\mathbf{w}, \mathbf{v})$ with fixed "reference" $\mathbf{v}$ [14]. In [14] $\mathbf{v}$ is kept unchanged during the whole optimization, whereas here on the contrary, $\mathbf{v}$ is updated after each one-dimensional optimization.

According to the latter point of view, the idea that the reference signal may contain *a priori* information on the separator re-enters, since the original "reference" is $\mathbf{v}_0$ and corresponds to the initialization point of algorithm **Alg.0**.

Note finally that for many sources, assumption A3 is valid: in the context of digital communications in particular, the cumulants are generally negative and can also often be assumed i.i.d. [18]. Consequently, Proposition 1 generally holds for digital communication signals. Nevertheless some signals such as those stemming from a continuous phase modulation (CPM) may be non i.i.d. [7]. An interesting question thus consists in knowing whether Proposition 1 remains true without Assumption A3. This issue will be addressed in simulations in Section VI, from which it will appear that the answer is likely to be positive. Another indication that Proposition 1 is likely to be true without assumption A3 is provided by the fact that the different contrast functions (both "reference-based" and kurtosis) have been proved in the general context of non i.i.d. signals with no condition on the signs of the cumulants.

## IV. CONVERGENCE ISSUES

This section is concerned with convergence issues of the proposed algorithms. First, we explain how a renormalization step should be introduced and we justify the reasons why it does not change the convergence properties of the algorithm. Then, the basic ideas underlying the validity of **Alg.0** are discussed. Finally, a connection with other optimization methods is made.

### A. Renormalization

It is known that the separating property of a separator is unaffected by a scaling factor, because of the unavoidable scaling ambiguity in BSS. Therefore, it is common in BSS to impose the unit-power constraint $\mathbb{E}\{|y(n)|^2\} = 1$. This can

be done by introducing a re-normalizing step in our algorithm, which yields algorithm **Alg.1**:

---

Algorithm **Alg.1**

- Initialize $\mathbf{v}_0$ and the corresponding reference signal $z_0(n) = \{\mathbf{v}_0\}\mathbf{x}(n)$.
- For $k = 0, 1, 2, \ldots, k_{\max} - 1$, repeat *(M0'-U)*:

*(M0')*   $\ast$ Set $\mathbf{d}_k = \nabla_1 \mathcal{I}(\mathbf{v}_k, \mathbf{v}_k)$,
  $\ast$ One-dimensional optimization:
  $\alpha_k = \arg\max_\alpha \mathcal{I}(\mathbf{v}_k + \alpha \mathbf{d}_k, \mathbf{v}_k)$.
  $\ast$ Set $\widetilde{\mathbf{v}}_{k+1} \leftarrow \mathbf{v}_k + \alpha_k \mathbf{d}_k$ and renormalize:
  $$\widetilde{\mathbf{v}}_{k+1} \leftarrow \frac{\widetilde{\mathbf{v}}_{k+1}}{(\mathbb{E}\{|\{\widetilde{\mathbf{v}}_{k+1}\}\mathbf{x}(n)|^2\})^{1/2}}$$

*(U)* Update: $\mathbf{v}_{k+1} \leftarrow \widetilde{\mathbf{v}}_{k+1}$

---

It is known that the unit-power constraint is equivalent to a unit-norm constraint on the separating or reference filter $\mathbf{v}$. Consequently, the points obtained after renormalization by **Alg.1** belong to the unit-sphere. In addition, it must be stressed that for $\lambda \neq 0$,

$$\mathcal{I}(\lambda\mathbf{w}, \mathbf{v}) = \mathcal{I}(\mathbf{w}, \mathbf{v}) \text{ and: } \mathcal{J}(\lambda\mathbf{w}) = \mathcal{J}(\mathbf{w}). \quad (7)$$

It hence follows that for any $\lambda \neq 0$, the gradient directions at $\mathbf{v}_k$ or at $\lambda\mathbf{v}_k$ are the same and more precisely:

$$\nabla_1 \mathcal{I}(\lambda\mathbf{w}, \mathbf{v}) = \frac{1}{\lambda}\nabla_1 \mathcal{I}(\mathbf{w}, \mathbf{v}) \text{ and: } \nabla\mathcal{J}(\lambda\mathbf{w}) = \frac{1}{\lambda}\nabla\mathcal{J}(\mathbf{w}). \quad (8)$$

Due to the scale invariance properties in Equations (7) and (8), it follows that the points (*i.e.* the separators) generated by algorithms **Alg.0** and **Alg.1** correspond to each other up to a scaling factor. This is illustrated in Figure 2 with a schematic representation of the successive points generated by **Alg.1** and **Alg.0**. In this figure, the criterion $\mathcal{J}$ has the same value all along each line passing through the origin.

It follows from the previous paragraph, that both algorithms **Alg.0** and **Alg.1** are theoretically equivalent and both are convergent according to Proposition 1. However, one can now understand from Figure 2 that a drift is observed when performing **Alg.0**: the power of the separator's output increases and may diverge unacceptably leading to a numerical overflow if a great number of iterations $k_{\max}$ is required by the algorithm. This phenomenon is avoided with the renormalizing step of **Alg.1**. This is also the reason why we did not consider **Alg.0** in simulations. Finally, it has been assumed in Proposition 1 that the sequence of points $(\mathbf{v}_k)_{k\in\mathbb{N}}$ is contained in a compact set. This condition is required in order to be able to state the convergence result, but it causes no difficulty in practice since one often considers FIR separators. In the latter case indeed, introducing a normalizing step ensures that the compactness assumption of Proposition 1 is satisfied.

### B. Validity of algorithm **Alg.0** and **Alg.1**

The complete proof of the convergence of **Alg.0** (and hence **Alg.1**) is provided in Appendix A. In this Section, we state and discuss the main point on which all the results are based. We systematically consider that the signals satisfy unit-power
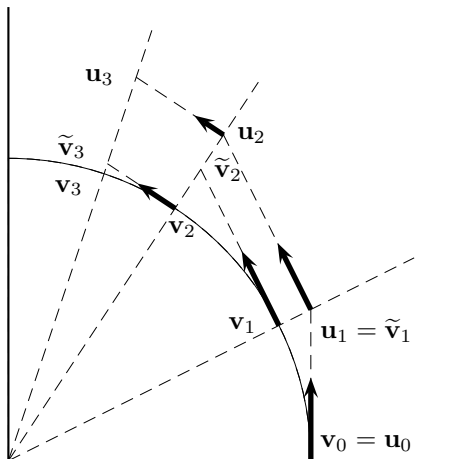
Fig. 2. Schematic comparison of the successive points obtained with **Alg.0** and **Alg.1**. For $k = 0, 1, 2, 3$, $\mathbf{u}_k$ is generated by **Alg.0**, $\widetilde{\mathbf{v}}_k$ is generated by **Alg.1** before renormalization, and $\mathbf{v}_k$ is the sequence generated by **Alg.1**.

constraint and we hence discuss the convergence of **Alg.1**. However, all the properties implicitly apply to **Alg.0**.

*1) Discussion and preliminary results:* The results in this paper mainly rely on the proposition which follows.

*Proposition 2:* We assume that assumption A3 and the constraints $\mathbb{E}\{|y(n)|^2\} = \mathbb{E}\{|z(n)|^2\} = 1$ hold. Then, if $|\mathrm{C}\{z\}| \leq |\mathcal{C}_z\{y\}|$, either one of the following situation holds (but not both):

(i) $|\mathcal{C}_z\{y\}| < |\mathrm{C}\{y\}|$

(ii) for all $i$ and $k$, $|g_i(p)|^2 = |t_i(p)|^2$, where $g_i(p)$ and $t_i(p)$ are defined in Equations (3)-(4).

As particular cases, we have the following implications:

$$|\mathrm{C}\{z\}| \leq |\mathcal{C}_z\{y\}| \quad \Rightarrow \quad |\mathcal{C}_z\{y\}| \leq |\mathrm{C}\{y\}| \quad (9)$$

$$|\mathrm{C}\{z\}| < |\mathcal{C}_z\{y\}| \quad \Rightarrow \quad |\mathcal{C}_z\{y\}| < |\mathrm{C}\{y\}| \quad (10)$$

The proof of the above result is relegated to Appendix B since the technical details are not required for a global understanding. However, the important ideas can now be understood. Remind that the ultimate goal is to extract one source and that, in our contrast function approach, this will be achieved by maximizing $\left|\frac{\mathrm{C}\{y\}}{\mathbb{E}\{|y|^2\}^2}\right|$: due to scaling ambiguity, this amounts to finding $y(n)$ which maximizes $|\mathrm{C}\{y\}|$ under the constraint $\mathbb{E}\{|y(n)|^2\} = 1$. The idea is to construct iteratively a signal sequence $y^{(k)}(n), k = 0, 1, \ldots$ of estimates of a source such that $|\mathrm{C}\{y^{(k)}\}|$ increases and converges to a maximum value (while keeping the constraint $\mathbb{E}\{|y^{(k)}(n)|^2\} = 1$). For this purpose, Algorithm **Alg.1** replaces the original criterion $|\mathrm{C}\{y\}|$ by the sequence of criteria $|\mathcal{C}_z\{y\}|$ where $z$ takes the successive values $y^{(k)}, k = 0, 1, \ldots$. The novel criteria are quadratic and can thus be efficiently optimized. Algorithm **Alg.1** finally constructs the sequence such that $|\mathcal{C}_{y^{(k-1)}}\{y^{(k)}\}| > |\mathrm{C}\{y^{(k-1)}\}|$. Proposition 2 then asserts that we have $|\mathrm{C}\{y^{(k)}\}| > |\mathcal{C}_{y^{(k-1)}}\{y^{(k)}\}| > |\mathrm{C}\{y^{(k-1)}\}|$ and hence we have indeed $|\mathrm{C}\{y^{(k)}\}| > |\mathrm{C}\{y^{(k-1)}\}|$: the criterion hence monotonically increases, which is the basis of our algorithm.

*2) Related optimization methods:* The algorithm **Alg.0** (or **Alg.1**) can be seen as iterative estimation algorithm of a source

signal. Starting from an initial "reference" $\mathbf{v}_0$ corresponding to a rough estimate of one MISO separating filter, it iteratively produces a sequence of better estimates $\mathbf{v}_k$ for the separating system. This point of view reveals the similarity of our method with Expectation-Maximization (EM) or Generalized Expectation-Maximization (GEM) methods [19]. An EM algorithm indeed maximizes a likelihood function $\mathcal{L}(\mathbf{v})$ by producing a set of successive estimates $\mathbf{v}_k$ obtained by iteratively maximizing a so-called Q-function :

$$\mathbf{v}_{k+1} = \arg\max_{\mathbf{v}} \mathcal{Q}(\mathbf{v}, \mathbf{v}_k) \quad (11)$$

In EM methods, the Q-function is obtained by introducing some hidden data in the model and taking the expectation of the complete log-likelihood.

In many contexts and specifically in our situation, it is difficult (or impossible) to identify any hidden variable. Fortunately, there exist generalizations of EM methods, which, depending on context, are referred to as Bound Optimization Algorithms [20], Surrogate Objective Functions [21] or MM (Maximization-Minimization or Minimization-Maximization) methods [22], [23]. In contrast to EM methods, MM methods do not require to identify a complete data set and take advantage of convexity and inequalities satisfied by the objective function.

The MM (and EM) methods, which are based on (11) exhibit interesting convergence properties under the condition that at each step, the Q-function satisfies:

$$\forall \mathbf{v}, \quad \mathcal{L}(\mathbf{v}) \geq \mathcal{Q}(\mathbf{v}, \mathbf{v}_k) \quad \text{and} \quad \mathcal{L}(\mathbf{v}_k) = \mathcal{Q}(\mathbf{v}_k, \mathbf{v}_k) \quad (12)$$

Such a condition can be proved in the EM context using Jensen's inequality. It follows from this condition that:

$$\mathcal{L}(\mathbf{v}_{k+1}) \geq \underbrace{\mathcal{L}(\mathbf{v}_{k+1}) - \mathcal{Q}(\mathbf{v}_{k+1}, \mathbf{v}_k)}_{\geq 0} + \mathcal{Q}(\mathbf{v}_{k+1}, \mathbf{v}_k)$$

$$\geq \mathcal{Q}(\mathbf{v}_{k+1}, \mathbf{v}_k)$$

$$\geq \mathcal{Q}(\mathbf{v}_k, \mathbf{v}_k) = \mathcal{L}(\mathbf{v}_k)$$

where the last inequality follows from (11). Consequently, the objective criterion monotonically increases, which results in the convergence of the algorithm.

One can see that algorithm **Alg.0** satisfies almost all properties of the MM maximization methods with $\mathcal{L}$ corresponding to $\mathcal{J}$ and $\mathcal{Q}$ corresponding to $\mathcal{I}$. However, there remain the following differences between **Alg.0** and the MM methods:

- Instead of (11), **Alg.0** performs a one-dimensional optimization and we are only able to ensure the weaker condition on $\mathbf{v}_{k+1}$: $\mathcal{Q}(\mathbf{v}_{k+1}, \mathbf{v}_k) \geq \mathcal{Q}(\mathbf{v}_k, \mathbf{v}_k)$. Such algorithms are referred to GEM in [19].

- In our method, the inequality (12) holds only for $\mathbf{v}$ such that $\mathcal{Q}(\mathbf{v}, \mathbf{v}_k) \geq \mathcal{Q}(\mathbf{v}_k, \mathbf{v}_k)$ (see Equations (9) and (10)). However, since one maximizes $\mathcal{Q}(\mathbf{v}, \mathbf{v}_k)$ with respect to $\mathbf{v}$, this does not alter the monotonic increase of the criterion.

## V. IMPROVEMENTS AND PRACTICAL IMPLEMENTATION

### A. Reference-based contrast functions and reference update

A natural idea in order to improve algorithm **Alg.0** or **Alg.1** consists in replacing the step *(M0)* or *(M0')* by a more efficient

operation: for instance, one can replace the optimization along a line in *(M0)* by a global optimization. The step *(M0)* would then become $\mathbf{v}_{k+1} \leftarrow \arg\max_{\mathbf{v}} \mathcal{I}(\mathbf{v}, \mathbf{v}_k)$. One can see that this algorithm corresponds to the "fixed-point" like algorithm proposed in [11], where the global optimization has been performed using an SVD decomposition. Alternatively, it can be seen from (11) that this corresponds to an EM (or MM) like algorithm. Considering the sensitivity of such an SVD-based optimization, it has been proposed to use a gradient like optimization or to use algorithm **Alg.1** [14], [15]. In the following algorithm, we propose to perform a given number of gradient optimization steps.

Algorithm **Alg.2**

---
- Initialize $\mathbf{v}_0$ and the corresponding reference signal $z_0(n) = \{\mathbf{v}_0\}\mathbf{x}(n)$.
- For $k = 0, 1, \ldots, k_{\max} - 1$, set $\widetilde{\mathbf{v}}_0 = \mathbf{v}_k$ and repeat *(M1'-U)*:

  *(M1')* For $l = 0, 1, 2, \ldots, l_{\max} - 1$ repeat *(M0')*:

  *(M0')* · Set $\widetilde{\mathbf{d}}_l = \nabla_1 \mathcal{I}(\widetilde{\mathbf{v}}_l, \mathbf{v}_k)$
  · One-dimensional optimization:
  $\widetilde{\alpha} = \arg\max_\alpha \mathcal{I}(\widetilde{\mathbf{v}}_l + \alpha\widetilde{\mathbf{d}}_l, \mathbf{v}_k)$.
  · Set: $\widetilde{\mathbf{v}}_{l+1} \leftarrow \widetilde{\mathbf{v}}_l + \widetilde{\alpha}\widetilde{\mathbf{d}}_l$ and renormalize:
  $$\widetilde{\mathbf{v}}_{l+1} \leftarrow \frac{\widetilde{\mathbf{v}}_{l+1}}{(\mathbb{E}\{|\{\widetilde{\mathbf{v}}_{l+1}\}\mathbf{x}(n)|^2\})^{1/2}}$$

  *(U)* Update: $\mathbf{v}_{k+1} \leftarrow \widetilde{\mathbf{v}}_{l_{\max}}$.

---

Note that a parameter $l_{\max}$ has been introduced in algorithm **Alg.2** in addition to $k_{\max}$. Depending on these two parameter values, **Alg.2** is intermediate between algorithm **Alg.1** ($l_{\max} = 1$) and a "fixed-point" like algorithm as proposed in [11] ($l_{\max}$ infinite). In addition, one can see that algorithm **Alg.2** belongs to the class of MM methods as described in Section IV-B2. It follows that **Alg.2** monotonically increases the criterion, which justifies its convergence. These arguments also explain the convergence of the "fixed-point" algorithm in [11].

### B. Implementation

We now describe how the method of algorithm **Alg.2** can be implemented in practice[2]. We consider FIR separators of given length $D$. The conditions under which such a separator exists can be found in [7], [11]. The separating filter $\mathbf{w}$ is then replaced by the following $1 \times QD$ vector which concatenates the vectors of the separator impulse response:

$$\underline{\mathbf{w}} \triangleq (\mathbf{w}(0), \mathbf{w}(1), \ldots, \mathbf{w}(D-1)) \tag{13}$$

Similarly, we define $\underline{\mathbf{v}} \triangleq (\mathbf{v}(0), \mathbf{v}(1), \ldots, \mathbf{v}(D-1))$ and the following $QD \times 1$ column vector:

$$\underline{\mathbf{x}}(n) \triangleq \left( \mathbf{x}(n)^T \quad \mathbf{x}(n-1)^T \quad \ldots \quad \mathbf{x}(n-D+1)^T \right)^T. \tag{14}$$

Now, let $\mathbf{R}$ be the covariance matrix of $\underline{\mathbf{x}}(n)$

$$\mathbf{R} \triangleq \mathbb{E}\{\underline{\mathbf{x}}(n)\underline{\mathbf{x}}(n)^H\} \tag{15}$$

[2]The `Matlab` code corresponding to our method is available at: http://www-public.it-sudparis.eu/~castella/toolbox/

and let $\mathbf{C}(\underline{\mathbf{v}})$ be the cumulant matrix defined component-wise by

$$(\mathbf{C}(\underline{\mathbf{v}}))_{i,j} \triangleq \mathrm{Cum}\{\underline{x}_i(n), \underline{x}_j^*(n), z(n), z^*(n)\} \tag{16}$$

where $\underline{x}_i(n), \underline{x}_j(n)$ are the components $i$ and $j$ of $\underline{\mathbf{x}}(n)$ respectively. Using these notations, it is shown in Appendix C, that the criterion $\mathcal{I}$ reads:

$$\mathcal{I}(\underline{\mathbf{w}}, \underline{\mathbf{v}}) = |\tilde{\mathcal{I}}(\underline{\mathbf{w}}, \underline{\mathbf{v}})|^2 \qquad \text{where:}$$
$$\tilde{\mathcal{I}}(\underline{\mathbf{w}}, \underline{\mathbf{v}}) = \frac{\underline{\mathbf{w}}\mathbf{C}(\underline{\mathbf{v}})\underline{\mathbf{w}}^H}{(\underline{\mathbf{v}}\mathbf{R}\underline{\mathbf{v}}^H)(\underline{\mathbf{w}}\mathbf{R}\underline{\mathbf{w}}^H)} \tag{17}$$

Then (see Appendix C), the gradient direction $\nabla_1 \mathcal{I}(\mathbf{v}, \mathbf{w})$ corresponds to the complex gradient vector $\underline{\mathbf{d}} = \frac{\partial \mathcal{I}}{\partial \underline{\mathbf{w}}^*}$ given below:

$$\frac{\partial \mathcal{I}}{\partial \underline{\mathbf{w}}^*} = \left(2\tilde{\mathcal{I}}(\underline{\mathbf{w}}, \underline{\mathbf{v}})\frac{\partial \tilde{\mathcal{I}}}{\partial \underline{\mathbf{w}}}\right)^* \qquad \text{with:} \tag{18}$$

$$\frac{\partial \tilde{\mathcal{I}}}{\partial \underline{\mathbf{w}}} = \frac{1}{\underline{\mathbf{v}}\mathbf{R}\underline{\mathbf{v}}^H}\left[\frac{\underline{\mathbf{w}}^*\mathbf{C}(\underline{\mathbf{v}})}{\underline{\mathbf{w}}\mathbf{R}\underline{\mathbf{w}}^H} - (\underline{\mathbf{w}}^*\mathbf{C}(\underline{\mathbf{v}})\underline{\mathbf{w}}^H)\frac{\underline{\mathbf{w}}^*\mathbf{R}}{(\underline{\mathbf{w}}\mathbf{R}\underline{\mathbf{w}}^H)^2}\right] \tag{19}$$

At step *(M0')* of algorithm **Alg.1** (respectively **Alg.2**), the vector $\underline{\mathbf{d}}_k$ (respectively $\widetilde{\mathbf{d}}_l$) is given by the above equations where $\underline{\mathbf{v}}$ and $\underline{\mathbf{w}}$ are set to $\underline{\mathbf{w}} = \underline{\mathbf{v}} = \underline{\mathbf{v}}_k$ (respectively $\underline{\mathbf{w}} = \widetilde{\underline{\mathbf{v}}}_l, \underline{\mathbf{v}} = \underline{\mathbf{v}}_k$). Because of the renormalization step, the multiplicative term $\underline{\mathbf{v}}\mathbf{R}\underline{\mathbf{v}}^H$ is equal to 1 and need not be computed in the algorithm.

Finally, according to Appendix C, the step-size of the one-dimensional optimization (that is the value of $\alpha_k$ in **Alg.1** or of $\widetilde{\alpha}$ in **Alg.2**) is a root of a second order polynomial $a_2\alpha^2 + a_1\alpha + a_0$. The coefficients $a_2, a_1, a_0$ are given in Appendix in Equations (29), (30), (31), where one should set $\underline{\mathbf{d}} = \underline{\mathbf{d}}_k$, $\underline{\mathbf{v}} = \underline{\mathbf{v}}_k$ and $\underline{\mathbf{w}} = \underline{\mathbf{v}}_k$ in **Alg.1** (or $\underline{\mathbf{w}} = \widetilde{\underline{\mathbf{v}}}_l$ in **Alg.2**). Checking which root yields the greatest value of the criterion yields the searched step-size.

One can notice that, similarly to "reference-based" contrasts in [11], $\mathcal{I}(\underline{\mathbf{w}}, \underline{\mathbf{v}}_k)$ depends quadratically on $\underline{\mathbf{w}}$. It follows that the one-dimensional optimization step in algorithms **Alg.1** and **Alg.2** is easier to perform than in a classical gradient ascent method.

## VI. SIMULATIONS

### A. Performance of the algorithms

*1) Validity and separation quality:*

*a) Simulation settings:* For different number of samples, $N = 3$ complex valued i.i.d. QAM4 sources have been generated taking values in $\{e^{i\pi/4}, e^{-i\pi/4}, e^{+i3\pi/4}, e^{-i3\pi/4}\}$ with equal probability $1/4$. They have been mixed by mixing filters with randomly driven coefficients, length $L = 3$, and $Q = 4$ sensors. The separating FIR separator has been searched with length $D = N(L-1) = 6$. An inverse of length $D$ exists generically for the FIR randomly chosen mixing system (see [7], [11] and references therein for details).

We tested the validity of algorithms **Alg.1** and **Alg.2**. A comparison has also been made with a gradient maximization of the criterion $\mathcal{J}$ (kurtosis) as in [8]. The latter has been programmed with an optimal step-size at each line-search. The

corresponding optimal step is obtained as a root of a fourth order polynomial, similarly to our algorithms. In addition a comparison with a gradient optimization of a reference based contrast function as in [14] is available through the the results where $k_{\max} = 1$ in **Alg.2**. The initialization of **Alg.1**/**Alg.2** has been set to $\underline{\mathbf{v}} = (0, \ldots, 0, 1, 0, \ldots, 0)$, where the value 1 is situated at a center position of the vector $\underline{\mathbf{v}}$ (or next to the center on the left if $\underline{\mathbf{v}}$ is of even length). With no additional assumption, there is no other possibility than choosing an arbitrary value for the initialization point. Due to the fact that the mixing system is random, the effect of the initial value is averaged in our experiment and similar average values should be obtained with any other initialization of the algorithms.

*b) Extraction of one source:* The separation results are reported in Table I where the average and median values of the mean square error (MSE) on the reconstructed source are given. The top row corresponds to a gradient maximization of the criterion $\mathcal{J}$ (kurtosis). The second row collects the results provided by **Alg.1**, which is the algorithm closest to a gradient on $\mathcal{J}$, as explained in Section III-C. The successive rows show the results provided by **Alg.2** with different values of the parameters $k_{\max}, l_{\max}$: a lower row indicates that the corresponding algorithm is farther from a gradient maximization of $\mathcal{J}$ and closer to a "reference-based" approach of source separation. Finally, last row gives the results for a gradient optimization of a "reference-based" contrast function.

From the first two rows of Table I, it can be observed that a gradient maximization of the kurtosis and algorithm **Alg.1** both yield similar values of the MSE. This confirms the validity of the method **Alg.1** and its convergence to a separating point. Additionally, an equally good (or even better) value of the MSE can be obtained by using algorithm **Alg.2**, which proves its validity.

*c) Extraction of all sources:* It has been illustrated that reference-based contrasts may be unsatisfying in a deflation scenario [14]. We hence tested here the behavior of our optimization algorithms in a deflation scenario. The results are provided in Table II and show that **Alg.2** behaves equally well as a kurtosis maximization.

| extracted source number | | 1st | 2nd | 3rd |
|---|---|---|---|---|
| average | kurtosis | 0.0029 | 0.0045 | 0.0079 |
| | **Alg.2** | 0.0026 | 0.0043 | 0.0076 |
| median | kurtosis | 0.0007 | 0.0034 | 0.0060 |
| | **Alg.2** | 0.0005 | 0.0033 | 0.0059 |

TABLE II
AVERAGE AND MEDIAN MSE FOR A DEFLATION EXTRACTION OF $N = 3$ QAM4 SOURCES. $T = 5000$ SAMPLES, $Q = 4$, $L = 3$. $k_{\max} = 50, l_{\max} = 20$. NUMBER OF MONTE-CARLO RUNS = 1000.

*2) Computational load:* On the one hand, methods relying on the kurtosis contrast function are known to provide good results. On the other hand, it has been shown that reference-based contrast functions yield methods which are significantly quicker. However, the result quality depends on the closeness of the reference to an actual source, and reference-based contrast functions hence require to be used with a so called "fixed-point" iteration [11].

We show here that algorithm **Alg.2** combines the advantages of both methods. It has been observed in the previous section that **Alg.1** and **Alg.2** yield a separation which is as good as the separation provided by a kurtosis maximization. Now, the three last columns in Table I clearly indicate that a good choice of $k_{\max}, l_{\max}$ leads to a significantly lower computational time. This is particularly striking with large number of samples, where the execution time is significantly reduced (see for example the values in bold in Table I).

To ensure a fair comparison, we checked the influence of the number of iterations: the number of iterations is given by $k_{\max}$ in a kurtosis gradient optimization, whereas the total number of iterations is given by the product $k_{\max} l_{\max}$ in the case of **Alg.2** (**Alg.1** is a particular case of **Alg.2** where $l_{\max} = 1$). The results corresponding to a fixed number of iterations are in Tables III, IV, V for respectively 1000, 5000 and 10000 available samples. The results corresponding to 1000 iterations have already been gathered in Table I. One can notice that algorithm **Alg.2** leaves a degree of freedom through the choice of $k_{\max}$ and $l_{\max}$. This is further investigated next.

*3) Complexity analysis:* The previous experimental results can be explained by some elements concerning the complexity of **Alg.2**. A careful inspection indeed reveals that the most time consuming steps in **Alg.2** are:

(i) the estimation of $\mathbf{C}(\underline{\mathbf{v}})$ defined in Equation (16),
(ii) the computation of the criterion value $\mathcal{I}(\widetilde{\mathbf{v}}_l, \mathbf{v}_k)$ and its gradient $\nabla_1 \mathcal{I}(\widetilde{\mathbf{v}}_l, \mathbf{v}_k)$,
(iii) the computation of the optimal step size, which is given by the polynomial roots of the numerator in (28).

On the one hand, step (i) involves the estimation of the cumulants in Equation (16). Its complexity is thus an increasing function of both the number of samples $T$ and the number of parameters provided by $Q$ and $D$. On the other hand, steps (ii) and (iii) do not involve any estimator from the set of sample data. Indeed, the criterion and its derivative are given by Equations (17)-(19) and the quantities necessary for computing the optimal steps are given in Equations (28) to (31): all these equations do not involve any estimator, but only the matrices $\mathbf{R}$ and $\mathbf{C}(\underline{\mathbf{v}})$ which are estimated and stored previously. Consequently, the complexity of steps (ii) and (iii) together consequently only depends on the number of parameters, that is on $Q$ and $D$. Let $\beta(T, Q, D)$ denote the complexity of step (i) and $\gamma(Q, D)$ denote the complexity of steps (ii) and (iii).

Now, it can be seen from the description of **Alg.2** that step (i) is repeated $k_{\max}$ times through the algorithm because the matrix $\mathbf{C}(\underline{\mathbf{v}})$ changes only at each update step. The overall complexity and computational load related to step (i) is thus $\mathcal{O}(k_{\max} \beta(T, Q, D))$. On the contrary, steps (i) and (ii) are repeated $k_{\max} l_{\max}$ times through the algorithm, which yields a complexity related to steps (ii) and (iii) of order $\mathcal{O}(k_{\max} l_{\max} \gamma(Q, D))$. It follows that the order of the complexity of **Alg.2** is:

$$\mathcal{O}\Big(k_{\max} \beta(T, Q, D) + k_{\max} l_{\max} \gamma(Q, D)\Big) \qquad (20)$$

where $\beta$ and $\gamma$ are increasing functions of their parameters. It is out of the scope of the paper to specify further $\beta$ and

| separation method | | | average MSE | | | | median MSE | | | | average execution time (s) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Number of samples | | | | Number of samples | | | | Number of samples | | | |
| parameters | $k_{\max}$ | $l_{\max}$ | 500 | 1000 | 5000 | 10000 | 500 | 1000 | 5000 | 10000 | 500 | 1000 | 5000 | 10000 |
| kurtosis | 1000 | - | **0.0489** | **0.0033** | **0.0028** | **0.0025** | 0.0007 | 0.0008 | 0.0006 | 0.0006 | **1.68** | **2.59** | **9.79** | **21.54** |
| **Alg.1** | 1000 | 1 | 0.0312 | 0.0034 | 0.0028 | 0.0025 | 0.0011 | 0.0009 | 0.0007 | 0.0006 | 1.29 | 2.48 | 15.98 | 26.02 |
| **Alg.2** | 200 | 5 | 0.0311 | 0.0033 | 0.0028 | 0.0025 | 0.0010 | 0.0009 | 0.0006 | 0.0006 | 0.52 | 0.78 | 3.48 | 5.40 |
| **Alg.2** | 100 | 10 | 0.0282 | 0.0025 | 0.0025 | 0.0024 | 0.0006 | 0.0005 | 0.0005 | 0.0006 | 0.42 | 0.56 | 1.92 | 2.82 |
| **Alg.2** | 50 | 20 | 0.0335 | 0.0027 | 0.0026 | 0.0025 | 0.0007 | 0.0006 | 0.0005 | 0.0006 | 0.37 | 0.45 | 1.13 | 1.54 |
| **Alg.2** | 25 | 40 | **0.0427** | **0.0029** | **0.0027** | **0.0025** | 0.0009 | 0.0007 | 0.0005 | 0.0006 | **0.34** | **0.39** | **0.74** | **0.90** |
| **Alg.2** | 10 | 100 | 0.0628 | 0.0054 | 0.0028 | 0.0026 | 0.0014 | 0.0010 | 0.0006 | 0.0006 | 0.32 | 0.35 | 0.50 | 0.51 |
| **Alg.2** | 8 | 125 | 0.0682 | 0.0063 | 0.0029 | 0.0026 | 0.0017 | 0.0011 | 0.0006 | 0.0007 | 0.32 | 0.34 | 0.47 | 0.46 |
| **Alg.2** | 5 | 200 | 0.1112 | 0.0100 | 0.0030 | 0.0027 | 0.0034 | 0.0014 | 0.0006 | 0.0006 | 0.31 | 0.34 | 0.42 | 0.38 |
| reference | 1 | 1000 | 0.5654 | 0.3567 | 0.1206 | 0.0870 | 0.5355 | 0.2764 | 0.0529 | 0.0297 | 0.31 | 0.33 | 0.36 | 0.28 |

TABLE I

MSE AND EXECUTION TIME FOR DIFFERENT SEPARATION METHODS. (QAM4 SOURCES, RANDOM MIXING FILTER WITH N=3 INPUTS, Q=4 SENSORS, AND LENGTH L=3. 1000 MONTE-CARLO REALIZATION)

| iterations | 12 | 50 | | | 100 | | | 500 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| method | kurt | kurt | **Alg.1** | **Alg.2** | kurt | **Alg.1** | **Alg.2** | kurt | **Alg.1** | **Alg.2** |
| param $k_{\max}$ | 12 | 50 | 50 | 2 | 100 | 100 | 4 | 500 | 500 | 4 |
| $l_{\max}$ | - | - | 1 | 25 | - | 1 | 25 | - | 1 | 125 |
| average MSE | 0.279 | 0.0266 | 0.0267 | 0.0261 | 0.0194 | 0.0177 | 0.0154 | 0.0078 | 0.0065 | 0.0055 |
| median MSE | 0.0199 | 0.0195 | 0.0199 | 0.0182 | 0.0120 | 0.0119 | 0.0108 | 0.0022 | 0.0022 | 0.0016 |
| average time (s) | 0.14 | 0.15 | 0.13 | 0.09 | 0.27 | 0.25 | 0.11 | 1.33 | 1.29 | 0.47 |

TABLE III

AVERAGE AND MEDIAN MSE FOR DIFFERENT VALUES OF $k_{\max}, l_{\max}$. NUMBER OF SAMPLES = 1000, NUMBER OF MONTE-CARLO RUNS = 1000.

| iterations | 12 | 50 | | | 100 | | | 500 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| method | kurt | kurt | **Alg.1** | **Alg.2** | kurt | **Alg.1** | **Alg.2** | kurt | **Alg.1** | **Alg.2** |
| param $k_{\max}$ | 12 | 50 | 50 | 25 | 100 | 100 | 25 | 500 | 500 | 25 |
| $l_{\max}$ | - | - | 1 | 2 | - | 1 | 4 | - | 1 | 20 |
| average MSE | 0.0251 | 0.0245 | 0.0244 | 0.0242 | 0.0151 | 0.0151 | 0.0151 | 0.0051 | 0.0052 | 0.0049 |
| median MSE | 0.0181 | 0.0177 | 0.0177 | 0.0175 | 0.0098 | 0.0098 | 0.0097 | 0.0019 | 0.0019 | 0.0018 |
| average time (s) | 0.66 | 0.68 | 0.99 | 0.54 | 1.20 | 1.87 | 0.55 | 5.14 | 8.34 | 0.62 |

TABLE IV

AVERAGE AND MEDIAN MSE FOR DIFFERENT VALUES OF $k_{\max}, l_{\max}$. NUMBER OF SAMPLES = 5000, NUMBER OF MONTE-CARLO RUNS = 1000.

| iterations | 12 | 50 | | | 100 | | | 500 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| method | kurt | kurt | **Alg.1** | **Alg.2** | kurt | **Alg.1** | **Alg.2** | kurt | **Alg.1** | **Alg.2** |
| param $k_{\max}$ | 12 | 50 | 50 | 25 | 100 | 100 | 20 | 500 | 500 | 10 |
| $l_{\max}$ | - | - | 1 | 2 | - | 1 | 5 | - | 1 | 50 |
| average MSE | 0.0250 | 0.0243 | 0.0243 | 0.0244 | 0.0156 | 0.0156 | 0.0157 | 0.0053 | 0.0054 | 0.0052 |
| median MSE | 0.0188 | 0.0184 | 0.0183 | 0.0184 | 0.0104 | 0.0103 | 0.0103 | 0.0021 | 0.0021 | 0.0021 |
| average time (s) | 1.59 | 1.61 | 1.92 | 1.03 | 3.18 | 3.76 | 0.88 | 15.48 | 18.54 | 0.64 |

TABLE V

AVERAGE AND MEDIAN MSE FOR DIFFERENT VALUES OF $k_{\max}, l_{\max}$. NUMBER OF SAMPLES = 10000, NUMBER OF MONTE-CARLO RUNS = 1000.

$\gamma$, but in our experiments, we observed that the computational load given by $\beta(T, Q, D)$ clearly increased with $T$ and that $\beta(T, Q, D)$ became much bigger than $\gamma(Q, D)$ for $T > 1000$.

The complexity of **Alg.1** (or equivalently **Alg.0**) is given by the above results with $l_{\max} = 1$, that is:

$$\mathcal{O}\Big(k_{\max}\big(\beta(T, Q, D) + \gamma(Q, D)\big)\Big) \qquad (21)$$

Finally, note that a gradient optimization of the kurtosis contrast function [8] has a complexity which is of the same order of magnitude as **Alg.1**: indeed at each iteration, such an algorithm performs steps which complexity roughly correspond to the steps (i),(ii) and (iii) described above. It follows that the execution time of a kurtosis maximization is proportional to $k_{\max}$ and is close to the execution time of **Alg.1**.

The above analysis confirms the experimental results and the fact that the computational load of **Alg.2** highly depends on the number $T$ of samples for large value of $k_{\max}$ (see the top four rows in Table I). On the contrary, when $l_{\max}$ is great and $k_{\max}$ small, the computational load is quite independent of $T$ (see the rows at bottom of Table I). Finally, for large number of samples, the term $\beta(T, Q, D)$ dominates and the complexity is directly proportional to $k_{\max}$. The analysis can be summed up by a few empirical rules to indicate how $k_{\max}$ and $l_{\max}$ should be chosen:

- increasing $k_{\max}$ generally ensures a better convergence,
- for large number of samples $T$ (approximately $T > 10000$), increasing $k_{\max}$ leads to a prohibitive computational load. $k_{\max}$ should be as small as possible (say

$k_{\max} < 10$), whereas it is possible to increase $l_{\max}$ to obtain better performance,

- for small number of samples ($T \approx 1000$), $k_{\max}$ should be great enough (approximately $k_{\max} > 20$) to avoid poor separation quality.

The main advantage of **Alg.2** consists in the possibility of adjusting the above parameters.

### B. Experimental study without assumption A3

In this Section, we provide experiments where Assumption A3 is no longer satisfied. This issue may be important since Assumption A3 is not necessarily satisfied in some practical applications.

*1) Different signs of the sources' cumulants:* In order to study the influence of the cumulants sign, we generated two zero-mean, unit variance, uniformly distributed, i.i.d. sources (cumulant value -1.2) and one unit-variance Laplace i.i.d. source (cumulant value +3). The successive sources have been retrieved using a deflation approach. The separation results and average MSE values are gathered in Table VI. The values indicate that the method seems still valid, although Assumption A3 is violated in this experiment.

| extracted source number | | 1st | 2nd | 3rd |
|---|---|---|---|---|
| average | kurtosis | 0.0253 | 0.0402 | 0.0694 |
| | **Alg.2** | 0.0218 | 0.0385 | 0.0772 |
| median | kurtosis | 0.0233 | 0.0162 | 0.0176 |
| | **Alg.2** | 0.0193 | 0.0160 | 0.0183 |

TABLE VI
AVERAGE AND MEDIAN MSE FOR A DEFLATION EXTRACTION OF $N = 3$ SOURCES: 2 UNIFORM'S, 1 DOUBLE-SIDED EXPONENTIAL. $T = 5000$ SAMPLES, $Q = 4$, $L = 3$. $k_{\max} = 5, l_{\max} = 200$. NUMBER OF MONTE-CARLO RUNS = 1000.

*2) Non i.i.d. sources: continuous phase modulation:* We considered Continuous Phase Modulation (CPM) source signals, which are of particular interest in a communication application. These sources are non i.i.d. and hence do not satisfy Assumption A3. However, the general scalar filtering ambiguity that generally remains when separating non i.i.d. sources has been characterized in the case of a kurtosis-based contrast function [7], [24], [25]. It is hence interesting to try in simulation our optimization method with CPM sources. A typical source separation result is given on Figure 3: it illustrates that our method seems to perform similarly to the kurtosis based method [7]. We performed simulations in conditions similar to the ones in [7]: we considered three CPM sources with modulation indices 0.75, 0.4 and 0.2 which have been mixed with the MIMO channel[3] in [7]. We also performed the symbol decoding similarly to [7]. The bit error rate (BER) is given in Table VII for different signal to noise ratios (SNR). Table VIII provides similar results for a randomly driven mixing filter. It appears that algorithm **Alg.2** provides good results at a much lower computational cost than with the method proposed in [7]. In passing, we can also note that the separation quality degrades during the deflation

[3]The parameters and the impulse response of this filter are available at `http://www-public.it-sudparis.eu/~castella/ChannelCPM.zip`

process. The separation quality of the third recovered source is indeed worse than the quality of the second, which is worse than the quality of the first one: this is due to accumulation errors. This drawback of deflation methods is well-known and can be compensated by solutions proposed previously, such as reinitialization [7].
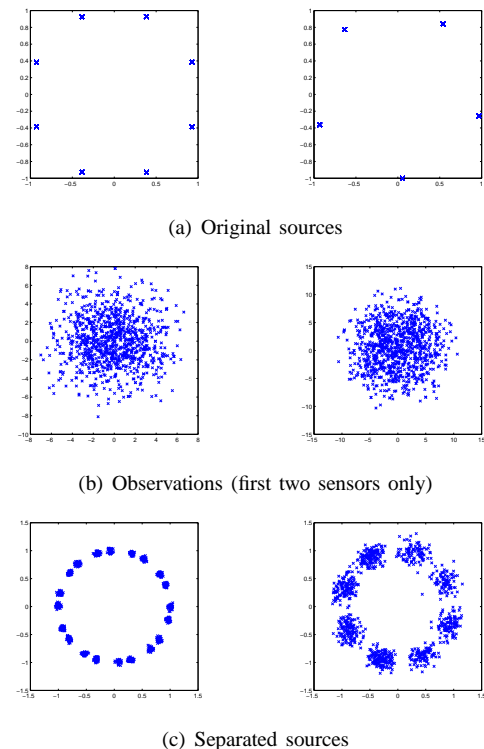


(a) Original sources

(b) Observations (first two sensors only)

(c) Separated sources

Fig. 3. Typical separation result of two CPM sources (N=2, Q=3, L=5).

### VII. CONCLUSION

In this paper, new methods have been proposed for the purpose of maximizing the kurtosis contrast function. They take advantage of the specificity of the kurtosis criterion and rely on the link between so-called "reference-based" contrast functions and the kurtosis. More precisely, we can say that our methods (**Alg.0**, **1** and **2**) perform several incomplete optimizations of a "reference-based" contrast: in other words, the reference is regularly updated before convergence of the optimization. As we showed, this establishes a connection with the EM and MM iterative optimization methods.

The advantage of our methods is twofold. First, we can theoretically prove the convergence to a stationary point: due to the separation property of all global maxima of the kurtosis, this implies in practice the convergence of the method to a separating filter. In addition, the method **Alg.2** allows one to adjust two iteration number parameters in order to improve the performance. In practice, an appropriate tuning of the number of iterations and reference updates allows one to significantly reduce the computational load. Simulations have indeed showed that algorithm **Alg.2** is particularly appealing, as it yields an impressive improvement in terms of computational speed for large number of samples. Finally, our results provide

| method | kurtosis | | | **Alg.2** ($l_{\max} = 125, k_{\max} = 8$) | | |
|---|---|---|---|---|---|---|
| execution time | $\approx 6.8$s | $\approx 6.6$s | $\approx 6.5$s | $\approx 0.92$s | $\approx 0.91$s | $\approx 0.89$s |
| SNR in dB | $\infty$ | 25 | 15 | $\infty$ | 25 | 15 |
| $1^{\text{st}}$ source: $\tau_1$ | 2.0e-6 | 4.2e-5 | 2.7e-3 | 3.0e-4 | 4.8e-5 | 1.7e-3 |
| $2^{\text{nd}}$ source: $\tau_2$ | 7.5e-4 | 1.2e-3 | 9.9e-3 | 4.2e-4 | 4.3e-4 | 3.2e-3 |
| $3^{\text{rd}}$ source: $\tau_3$ | 6.0e-3 | 6.2e-3 | 3.1e-2 | 4.3e-3 | 4.6e-3 | 1.7e-2 |

TABLE VII

ESTIMATION OF THE BER FOR A FIXED $4 \times 3$ MIXING FILTER AND FOR DIFFERENT VALUES OF THE SNR. THE NUMBER OF MONTE-CARLO RUNS WAS 1000 (SEE [7] FOR A COMPARISON).

| method | kurtosis | | | **Alg.2** ($l_{\max} = 125, k_{\max} = 8$) | | | **Alg.2** ($l_{\max} = 40, k_{\max} = 25$) | | |
|---|---|---|---|---|---|---|---|---|---|
| execution time | $\approx 5.0$s | $\approx 5.1$s | $\approx 8.8$s | $\approx 0.70$s | $\approx 0.72$s | $\approx 1.1$s | $\approx 0.77$s | $\approx 0.79$s | $\approx 1.2$s |
| SNR in dB | $\infty$ | 25 | 15 | $\infty$ | 25 | 15 | $\infty$ | 25 | 15 |
| $1^{\text{st}}$ source: $\tau_1$ | 6.8e-4 | 1.3e-3 | 1.7e-2 | 2.0e-3 | 2.5e-3 | 1.7e-2 | 5.9e-4 | 1.6e-3 | 1.8e-2 |
| $2^{\text{nd}}$ source: $\tau_2$ | 2.4e-3 | 3.8e-3 | 3.4e-2 | 8.1e-3 | 7.9e-3 | 3.1e-2 | 3.9e-3 | 4.2e-3 | 3.2e-2 |
| $3^{\text{rd}}$ source: $\tau_3$ | 7.0e-3 | 9.6e-3 | 5.9e-2 | 1.3e-2 | 1.8e-2 | 5.5e-2 | 1.0e-2 | 1.4e-2 | 5.9e-2 |

TABLE VIII

ESTIMATION OF THE BER FOR A $4 \times 3$ MIXING FILTER WITH RANDOM COEFFICIENTS AND FOR DIFFERENT VALUES OF THE SNR. THE NUMBER OF MONTE-CARLO RUNS WAS 1000 (SEE [7] FOR A COMPARISON).

a justification for the "fixed-point" iterations introduced in the context of "reference" contrast functions.

# APPENDIX A
## PROOF OF PROPOSITION 1

*Proof:* Let $\Gamma = \{\mathbf{v} \,|\, \nabla \mathcal{J}(\mathbf{v}) = 0\}$ be the set of stationary points of $\mathcal{J}$. Following the general framework of global convergence analysis [26], [27], we define the algorithm **Alg.0** as a point to set mapping denoted here by $\mathcal{A}_0$. The sequence of points $(\mathbf{v}_k)_{k \in \mathbb{N}}$ is generated such that it satisfies $\mathbf{v}_{k+1} \in \mathcal{A}_0(\mathbf{v}_k)$. More precisely, the algorithm $\mathcal{A}_0$ can be decomposed in $\mathcal{A}_0 = \mathcal{S}_0 \mathcal{G}_0$. Here $\mathcal{G}_0 : \mathbf{v} \mapsto (\mathbf{v}, -\mathbf{d})$ with $\mathbf{d} = \nabla_1 \mathcal{I}(\mathbf{v}, \mathbf{v})$ and $\mathcal{S}_0$ is a line search along the direction $-\mathbf{d}$. It is proved in [26, p.210] that the line search is a closed mapping at points where $\mathbf{d} \neq 0$, from which follows the closedness of the algorithm $\mathcal{A}_0$ outside $\Gamma$.

Finally, $-\mathcal{J}$ is a descent function. Consider indeed a separating filter $\mathbf{v}$, set $\mathbf{d} = \nabla_1 \mathcal{I}(\mathbf{v}, \mathbf{v})$ and define $\mathbf{w} = \mathbf{v} + \alpha \mathbf{d}$ where $\alpha \geq 0$ maximizes $\mathcal{I}(\mathbf{v} + \alpha \mathbf{d}, \mathbf{v})$. Then:

$$\mathcal{J}(\mathbf{v}) \triangleq \mathcal{I}(\mathbf{v}, \mathbf{v}) \leq \max_{\alpha \geq 0} \mathcal{I}(\mathbf{v} + \alpha \mathbf{d}, \mathbf{v}) = \mathcal{I}(\mathbf{w}, \mathbf{v}) \quad (22)$$

$$\leq \mathcal{I}(\mathbf{w}, \mathbf{w}) \triangleq \mathcal{J}(\mathbf{w}) \quad (23)$$

Inequality (23) follows from (9) where $y(n) = \frac{\{\mathbf{w}\}x(n)}{(\mathbb{E}\{|\{\mathbf{w}\}x(n)|^2\})^{1/2}}$ and $z(n) = \frac{\{\mathbf{v}\}x(n)}{(\mathbb{E}\{|\{\mathbf{v}\}x(n)|^2\})^{1/2}}$. The definition of a descent function further requires that for $\mathbf{v} \notin \Gamma$, the strict inequality $\mathcal{J}(\mathbf{v}) < \mathcal{J}(\mathbf{w})$ holds. Assume precisely $\mathbf{v} \notin \Gamma$, that is $\nabla \mathcal{J}(\mathbf{v}) \neq 0$. Since $\nabla_1 \mathcal{I}(v, v) = \frac{1}{2} \nabla \mathcal{I}(v, v)$, one can see that $\mathbf{d} = \nabla_1 \mathcal{I}(v, v) \neq 0$. Then, from the definition of $\nabla_1$, the function $\alpha \mapsto \mathcal{I}(\mathbf{v} + \alpha \mathbf{d}, \mathbf{v})$ is strictly increasing in a neighborhood of $\alpha = 0$, which implies that the inequality (22) is a strict one and thus $\mathcal{J}(\mathbf{v}) < \mathcal{J}(\mathbf{w})$ as required.

Proposition 1 then follows from Zangwill's global convergence theorem [26, p.187]. ∎

# APPENDIX B
## PROOF OF PROPOSITION 2

### A. Lemma

For the sake of clarity, we first prove an independent lemma in the real valued case. Here, $\mathbf{x}$, $\mathbf{y}$ denote vectors with respective components $x_i$, $y_i$. Let $\ell^1$ denote the set of absolutely summable sequences and let $(\alpha_i)_{i \in \mathbb{Z}}$ be a bounded and positive sequence. Let us define $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i \in \mathbb{Z}} \alpha_i x_i y_i$ and $\|x\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$.

*Lemma 1:* Let $\mathbf{x}$ and $\mathbf{y}$ be vectors in $\ell^1$ with non-negative components ($\forall i \in \mathbb{Z}, x_i \geq 0, y_i \geq 0$). Assume in addition that $\sum_{i \in \mathbb{Z}} x_i = \sum_{i \in \mathbb{Z}} y_i = 1$. If $\|\mathbf{x}\|^2 \leq \langle \mathbf{x}, \mathbf{y} \rangle$, then:

$$\mathbf{x} = \mathbf{y} \quad \Leftrightarrow \quad \|\mathbf{y}\|^2 \leq \langle \mathbf{x}, \mathbf{y} \rangle \quad (24)$$

*Proof:* First note that for any sequence in $\ell^1$ the sums $\sum_i x_i^2, \sum_i \alpha_i |x_i|$ and $\sum_i \alpha_i x_i^2$ converge and that $\langle ., . \rangle$ defines a scalar product. Then, obviously, $\mathbf{x} = \mathbf{y}$ implies $\|\mathbf{y}\|^2 \leq \langle \mathbf{x}, \mathbf{y} \rangle$.

Conversely, assume that $\|\mathbf{y}\|^2 \leq \langle \mathbf{x}, \mathbf{y} \rangle$. Noting that we have also assumed $\|\mathbf{x}\|^2 \leq \langle \mathbf{x}, \mathbf{y} \rangle$ and combining these assumptions with the Cauchy-Schwarz inequality, we obtain $\langle \mathbf{x}, \mathbf{y} \rangle^2 \leq \|\mathbf{x}\|^2 \|\mathbf{y}\|^2 \leq \langle \mathbf{x}, \mathbf{y} \rangle^2$ and thus, because the components of $\mathbf{x}$ and $\mathbf{y}$ are non-negative, $\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x}\|\|\mathbf{y}\|$. The vectors $\mathbf{x}$ and $\mathbf{y}$ are hence collinear and the condition $\sum_{i \in \mathbb{Z}} x_i = \sum_{i \in \mathbb{Z}} y_i = 1$ shows that in fact $\mathbf{x} = \mathbf{y}$. ∎

The above lemma will be used in the following corollary form:

*Corollary 1:* Let $\mathbf{x}$ and $\mathbf{y}$ be vectors in $\ell^1$ with non-negative components ($\forall i \in \mathbb{Z}, x_i \geq 0, y_i \geq 0$). Assume in addition that $\sum_{i \in \mathbb{Z}} x_i = \sum_{i \in \mathbb{Z}} y_i = 1$. If $\|\mathbf{x}\|^2 \leq \langle \mathbf{x}, \mathbf{y} \rangle$, then either one of the following conditions holds (but not both simultaneously):

(i) $\langle \mathbf{x}, \mathbf{y} \rangle < \|\mathbf{y}\|^2$

(ii) $\mathbf{x} = \mathbf{y}$

### B. Proof of Proposition 2

We are now in position to prove Proposition 2.

*Proof:* Let us define the following scalar product $\langle \mathbf{x}, \mathbf{y} \rangle_\kappa \triangleq \sum_{i=1}^{N} |\kappa_i| \sum_{p \in \mathbb{Z}} x_i(p) y_i(p)^*$ and denote by $\mathbf{g}^2$ and

$\mathbf{t}^2$ the vectors $\mathbf{g}^2 \triangleq (|g_i(p)|^2)_{i,k}$ and $\mathbf{t}^2 \triangleq (|t_i(p)|^2)_{i,k}$. Using the independence of the sources, the above notations, the relations (2)-(3), and assumption A3, we can write:

$$|\mathcal{C}_z\{y\}| = \langle \mathbf{g}^2, \mathbf{t}^2 \rangle_\kappa, \qquad |C\{y\}| = \langle \mathbf{g}^2, \mathbf{g}^2 \rangle_\kappa = \|\mathbf{g}^2\|_\kappa^2,$$
$$\text{and:} \quad |C\{z\}| = \langle \mathbf{t}^2, \mathbf{t}^2 \rangle_\kappa = \|\mathbf{t}^2\|_\kappa^2$$

Now, since the components of $\mathbf{g}^2$ and $\mathbf{t}^2$ are non-negative and satisfy $\sum_{i=1}^N \sum_{p \in \mathbb{Z}} |g_i(p)|^2 = \sum_{i=1}^N \sum_{p \in \mathbb{Z}} |t_i(p)|^2 = 1$ because of the constraint $\mathbb{E}\{|y(n)|^2\} = \mathbb{E}\{|z(n)|^2\} = 1$, the conditions of Corollary 1 hold. Hence, we deduce that if $|C\{z\}| \leq |\mathcal{C}_z\{y\}|$, either $\langle \mathbf{g}^2, \mathbf{t}^2 \rangle_\kappa < \|\mathbf{g}^2\|_\kappa^2$ or $\mathbf{g}^2 = \mathbf{t}^2$ holds (but not both): these are respectively conditions (i) and (ii) of Proposition 2.

(9) follows from the fact that $|\mathcal{C}_z\{y\}| \leq |C\{y\}|$ holds in both cases (i) and (ii) of the first part of the proposition. (10) follows from the fact that if there is a strict inequality $|\mathcal{C}_z\{y\}| < |C\{y\}|$, (ii) cannot hold and we have (i) necessarily. ∎

# APPENDIX C
## CRITERION DERIVATIVE AND OPTIMAL STEP

We here justify the results summed up in Section V-B. Remind the definition of $\underline{\mathbf{w}}, \underline{\mathbf{v}}$ in Equation (13) and the definition of $\underline{\mathbf{x}}(n)$ in Equation (14). It is then straightforward to see that the processed output and the "reference" signal defined in Equations (3) and (4) can be written as:

$$y(n) = \underline{\mathbf{w}}\,\underline{\mathbf{x}}(n) \qquad \text{and:} \qquad z(n) = \underline{\mathbf{v}}\,\underline{\mathbf{x}}(n).$$

Now, by construction, the matrices $\mathbf{R}$ and $\mathbf{C}(\underline{\mathbf{v}})$ in Equations (15), (16) are hermitian and we have:

$$\mathcal{C}_z\{y(n)\} = \underline{\mathbf{w}}\mathbf{C}(\underline{\mathbf{v}})\underline{\mathbf{w}}^H \tag{25}$$
$$\mathbb{E}\{|y(n)|^2\} = \underline{\mathbf{w}}\mathbf{R}\underline{\mathbf{w}}^H \qquad \text{and:} \ \mathbb{E}\{|z(n)|^2\} = \underline{\mathbf{v}}\mathbf{R}\underline{\mathbf{v}}^H \tag{26}$$

Equation (17) then easily follows from its definition in (5).

The complex derivative operator is defined by $\frac{\partial \cdot}{\partial z} \triangleq \frac{1}{2}\left(\frac{\partial \cdot}{\partial z_\Re} - \iota \frac{\partial \cdot}{\partial z_\Im}\right)$ and the complex conjugate derivative operator by $\frac{\partial \cdot}{\partial z^*} \triangleq \frac{1}{2}\left(\frac{\partial \cdot}{\partial z_\Re} + \iota \frac{\partial \cdot}{\partial z_\Im}\right)$, where $z_\Re, z_\Im$ are the complex and imaginary part of $z$. For any real-valued quantity, we have $\frac{\partial \cdot}{\partial z^*} = \left(\frac{\partial \cdot}{\partial z}\right)^*$. It follows that the gradient direction in (18) is indeed $\underline{\mathbf{d}} \triangleq \frac{\partial \mathcal{I}}{\partial \underline{\mathbf{w}}^*} = \left(\frac{\partial \mathcal{I}}{\partial \underline{\mathbf{w}}}\right)^* = \left(2\tilde{\mathcal{I}}(\underline{\mathbf{w}}, \underline{\mathbf{v}}) \frac{\partial \tilde{\mathcal{I}}}{\partial \underline{\mathbf{w}}}\right)^*$. One can then verify the following derivation rules (see [28] for more details):

$$\frac{\partial \underline{\mathbf{w}}\mathbf{R}\underline{\mathbf{w}}^H}{\partial \underline{\mathbf{w}}} = \underline{\mathbf{w}}^*\mathbf{R} \qquad \text{and:} \qquad \frac{\partial \underline{\mathbf{w}}\mathbf{C}(\underline{\mathbf{v}})\underline{\mathbf{w}}^H}{\partial \underline{\mathbf{w}}} = \underline{\mathbf{w}}^*\mathbf{C}(\underline{\mathbf{v}}) \tag{27}$$

and Equation (19) then follows.

Now, to derive the optimal step size at each one-dimensional optimization of the algorithms note that from (17), $\frac{\partial \mathcal{I}(\underline{\mathbf{w}}+\mu\underline{\mathbf{d}}, \underline{\mathbf{v}})}{\partial \mu} = 2\tilde{\mathcal{I}}(\underline{\mathbf{w}} + \mu\underline{\mathbf{d}}, \underline{\mathbf{v}}) \frac{\partial \tilde{\mathcal{I}}(\underline{\mathbf{w}}+\mu\underline{\mathbf{d}}, \underline{\mathbf{v}})}{\partial \mu}$. Since the algorithms searches for a maximum point and since $\mathcal{I}$ is positive, the optimal step should be such that $\frac{\partial \tilde{\mathcal{I}}(\underline{\mathbf{w}}+\mu\underline{\mathbf{d}}, \underline{\mathbf{v}})}{\partial \mu} = 0$.

Now, we have:

$$\tilde{\mathcal{I}}(\underline{\mathbf{w}} + \mu\underline{\mathbf{d}}, \underline{\mathbf{v}}) = \frac{1}{\underline{\mathbf{v}}\mathbf{R}\underline{\mathbf{v}}^H}\left[\frac{(\underline{\mathbf{w}} + \mu\underline{\mathbf{d}})\mathbf{C}(\underline{\mathbf{v}})(\underline{\mathbf{w}} + \mu\underline{\mathbf{d}})^H}{((\underline{\mathbf{w}} + \mu\underline{\mathbf{d}})\mathbf{R}(\underline{\mathbf{w}} + \mu\underline{\mathbf{d}})^H)}\right] =$$
$$\frac{1}{\underline{\mathbf{v}}\mathbf{R}\underline{\mathbf{v}}^H}\left[\frac{\mu^2(\underline{\mathbf{d}}\mathbf{C}(\underline{\mathbf{v}})\underline{\mathbf{d}}^H) + 2\mu\Re[\underline{\mathbf{w}}\mathbf{C}(\underline{\mathbf{v}})\underline{\mathbf{d}}^H] + \underline{\mathbf{w}}\mathbf{C}(\underline{\mathbf{v}})\underline{\mathbf{w}}^H}{\mu^2(\underline{\mathbf{d}}\mathbf{R}\underline{\mathbf{d}}^H) + 2\mu\Re[\underline{\mathbf{w}}\mathbf{R}\underline{\mathbf{d}}^H] + \underline{\mathbf{w}}\mathbf{R}\underline{\mathbf{w}}^H}\right]$$

and thus:

$$\frac{\partial \tilde{\mathcal{I}}(\underline{\mathbf{w}} + \mu\underline{\mathbf{d}}, \underline{\mathbf{v}})}{\partial \mu} =$$
$$\frac{2}{\underline{\mathbf{v}}\mathbf{R}\underline{\mathbf{v}}^H}\left[\frac{a_2\mu^2 + a_1\mu + a_0}{(\mu^2(\underline{\mathbf{d}}\mathbf{R}\underline{\mathbf{d}}^H) + 2\mu\Re[\underline{\mathbf{w}}\mathbf{R}\underline{\mathbf{d}}^H] + \underline{\mathbf{w}}\mathbf{R}\underline{\mathbf{w}}^H)^2}\right] \tag{28}$$

where:

$$a_2 = \underline{\mathbf{d}}\mathbf{C}(\underline{\mathbf{v}})\underline{\mathbf{d}}^H\Re[\underline{\mathbf{w}}\mathbf{R}\underline{\mathbf{d}}^H] - \Re[\underline{\mathbf{w}}\mathbf{C}(\underline{\mathbf{v}})\underline{\mathbf{d}}^H]\underline{\mathbf{d}}\mathbf{R}\underline{\mathbf{d}}^H \tag{29}$$
$$a_1 = \underline{\mathbf{d}}\mathbf{C}(\underline{\mathbf{v}})\underline{\mathbf{d}}^H\underline{\mathbf{w}}\mathbf{R}\underline{\mathbf{w}}^H - \underline{\mathbf{w}}\mathbf{C}(\underline{\mathbf{v}})\underline{\mathbf{w}}^H\underline{\mathbf{d}}\mathbf{R}\underline{\mathbf{d}}^H \tag{30}$$
$$a_0 = \Re[\underline{\mathbf{w}}\mathbf{C}(\underline{\mathbf{v}})\underline{\mathbf{d}}^H]\underline{\mathbf{w}}\mathbf{R}\underline{\mathbf{w}}^H - \underline{\mathbf{w}}\mathbf{C}(\underline{\mathbf{v}})\underline{\mathbf{w}}^H\Re[\underline{\mathbf{w}}\mathbf{R}\underline{\mathbf{d}}^H] \tag{31}$$

If follows that the optimal step is such that the numerator of (28) vanishes.

## REFERENCES

[1] F. Benedetto, G. Giunta, and L. Vandendorpe, "A blind equalization algorithm based on minimization of normalized variance for DS/CDMA communications," *IEEE Trans. on Vehicular Technology*, vol. 57, no. 6, pp. 3453–3461, November 2008.

[2] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation, Independent Component Analysis and Applications*. Academic Press, 2010.

[3] P. Comon, "Contrasts for multichannel blind deconvolution," *IEEE Signal Processing Letters*, vol. 3, no. 7, pp. 209–211, July 1996.

[4] P. Comon and E. Moreau, "Blind MIMO equalization and joint-diagonalization criteria," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, USA, May 2001.

[5] E. Moreau and J.-C. Pesquet, "Generalized contrasts for multichanel blind deconvolution of linear systems," *IEEE Signal Processing Letters*, vol. 4, no. 6, pp. 182–183, June 1997.

[6] M. Castella, J.-C. Pesquet, and A. P. Petropulu, "A family of frequency- and time-domain contrasts for blind separation of convolutive mixtures of temporally dependent signals," *IEEE Trans. Signal Processing*, vol. 53, no. 1, pp. 107–120, January 2005.

[7] M. Castella, P. Bianchi, A. Chevreuil, and J.-C. Pesquet, "A blind source separation framework for detecting CPM sources mixed by a convolutive MIMO filter," *Signal Processing*, vol. 86, no. 8, pp. 1950–1967, August 2006.

[8] C. Simon, P. Loubaton, and C. Jutten, "Separation of a class of convolutive mixtures: a contrast function approach," *Signal Processing*, vol. 4, no. 81, pp. 883–887, 2001.

[9] J. K. Tugnait, "Identification and deconvolution of multichannel linear non-gaussian processes using higher order statistics and inverse filter criteria," *IEEE Trans. Signal Processing*, vol. 45, no. 3, pp. 658–672, March 1997.

[10] D. N. Godard, "Self-recovering equalization and carrier tracking in two-dimensional data communication systems," *IEEE Trans. on Communications*, vol. 28, no. 11, pp. 1867–1875, November 1980.

[11] M. Castella, S. Rhioui, E. Moreau, and J.-C. Pesquet, "Quadratic higher-order criteria for iterative blind separation of a MIMO convolutive mixture of sources," *IEEE Trans. Signal Processing*, vol. 55, no. 1, pp. 218–232, January 2007.

[12] R. Dubroca, C. De Luigi, M. Castella, and E. Moreau, "A general algebraic algorithm for blind extraction of one source in a MIMO convolutive mixture," *IEEE Trans. Signal Processing*, vol. 58, no. 5, pp. 2484–2493, May 2010.

[13] M. Kawamoto, K. Kohno, and Y. Inouye, "Eigenvector algorithms incorporated with reference systems for solving blind deconvolution of MIMO-IIR linear systems," *IEEE Signal Processing Letters*, vol. 14, no. 12, pp. 996–999, December 2007.

[14] M. Castella and E. Moreau, "A new optimization method for reference-based quadratic contrast functions in a deflation scenario," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, R.O.C., April 2009, pp. 3161–3164.

[15] ——, "A new method for kurtosis maximization and source separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, TX, USA, March 2010, pp. 2670–2673.

[16] N. Delfosse and P. Loubaton, "Adaptive blind separation of independent sources: a deflation approach," *Signal Processing*, vol. 45, pp. 59–83, 1995.

[17] P. Loubaton and P. Regalia, "Blind deconvolution of multivariate signals: a deflation approach," in *Proceedings of ICC*, Geneva, Switzerland, May 1993, pp. 1160–1164.

[18] H. Mathis, "On the kurtosis of digitally modulated signals with timing offsets," in *IEEE Third Workshop on Signal Processing Advances in Wireless Communications (SPAWC'01)*, Taoyuan, Taiwan, ROC, March 2001, pp. 86–89.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[20] M. A. T. Figueiredo and R. D. Nowak, "A bound optimization approach to wavelet-based image deconvolution," in *IEEE Int. Conf. on Image Processing*, Genoa, Italy, 2005.

[21] K. Lange, D. R. Hunter, and I. Yang, "Optimization transfer using surrogate objective functions," *Journal of Computational and Graphical Statistics*, vol. 9, no. 1-20, 2000.

[22] D. R. Hunter and K. Lange, "A Tutorial on MM Algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, February 2004.

[23] M. A. T. Figueiredo, J. M. Bioucas-Dias, and R. D. Nowak, "Majorization-minimization algorithms for wavelet-based image restoration," *IEEE Trans. on Image Processing*, vol. 16, no. 12, pp. 2980–2991, December 2007.

[24] P. Bianchi, P. Loubaton, and F. Sirven, "On the blind equalization of continuous phase modulation using a constant modulus criterion," in *Proc. SPAWC*, Roma, Italy, 2003.

[25] P. Bianchi and P. Loubaton, "On the blind equalization of continuous phase modulated signals using the constant modulus criterion," *IEEE Trans. Signal Processing*, vol. 55, no. 3, pp. 1047–1061, March 2007.

[26] D. G. Luenberger, *Linear and nonlinear programming*, 2nd ed. Addison-Wesley, 1984.

[27] W. I. Zangwill, *Nonlinear Programming: A Unified Approach*. Prentice-Hall, Inc. Englewood Cliffs, N.J., 1969.

[28] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," oct 2008, version 20081110. [Online]. Available: http://www2.imm.dtu.dk/pubdb/p.php?3274

**Eric Moreau** was born in Lille, France. He graduated from the "Ecole Nationale Supérieure des Arts et Métiers" (ENSAM), Paris, France, in 1989 and received the "Agrégation de Physique" degree from the "Ecole Normale Supérieure de Cachan" in 1990. He received the DEA degree in 1991 and the Ph.D. degree in 1995, both in the field of signal processing and from the University of Paris-Sud, France.

From 1995 to 2001, he was assistant professor within the Telecommunications Department of the Engineering School "Institut des Sciences de l'Ingénieur de Toulon et du Var" (ISITV), La Valette, France. He is currently a Professor with the University of Toulon, France. His main research interests are in statistical signal processing using high-order statistics, source separation and tensors decomposition.

**Marc Castella** was born in Courbevoie, France. In 2000, he received the "Agrégation" degree in the field of applied physics and in 2001 he received the M.Sc. degree in electrical engineering, from both the "École Normale Supérieure de Cachan" and the "Université Paris-Sud, Orsay" (France). He obtained his Ph.D. degree from "Université de Marne-la-Vallée" (France) in 2004. Since then, he has been "Maître de Conférence" in Evry (France) at "Télécom SudParis". He is also a member of the UMR-CNRS 5157 research team SAMOVAR. His research activities have focused on the problem blind source separation in general.